This repository contains a tool developed to facilitate the calculation of biologically informed Polygenic Risk Score (ePRS).

The original method for ePRS calculation was published by Silveira Lab in 20171. The details of the method provided below are based on one of the ePRS, specifically the 5-HTT ePRS, described in de Lima et al 20202.

Generation of any ePRS starts from defining the critical determinant of the biological functionality that represents the research question. In this example we were interested in genes that were co-expressed with the serotonin transporter gene (5-HTT, GRCh37.p13/ENSG00000108576) in amygdala, during early life (fetal life/infancy).

**The Amygdala 5HTT ePRS**

The expression-based polygenic risk score was created considering genes co-expressed with the serotonin transporter gene (5-HTT-ePRS) in amygdala, according to the protocol previously described by Silveira et al(2017)1 and Miguel et al(2019)3.

Online resources/databases used for the expression-based polygenic risk score:

(A) GeneNetwork (<http://genenetwork.org>);

(B) BrainSpan (<http://www.brainspan.org>);

(C) NCBI Variation Viewer (<https://www.ncbi.nlm.nih.gov/variation/view>);

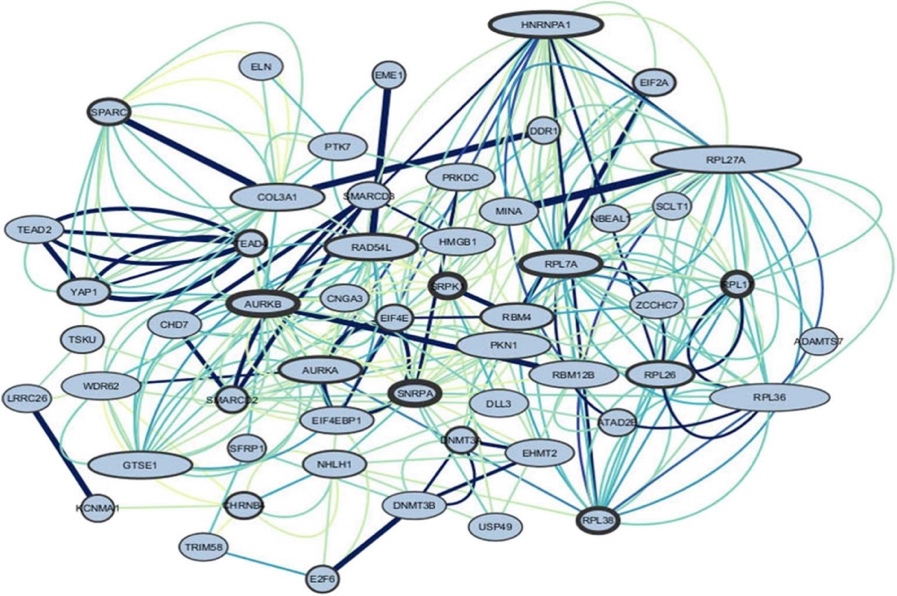
(D) The Genotype-Tissue Expression (GTEx) (<https://gtexportal.org/home/>).

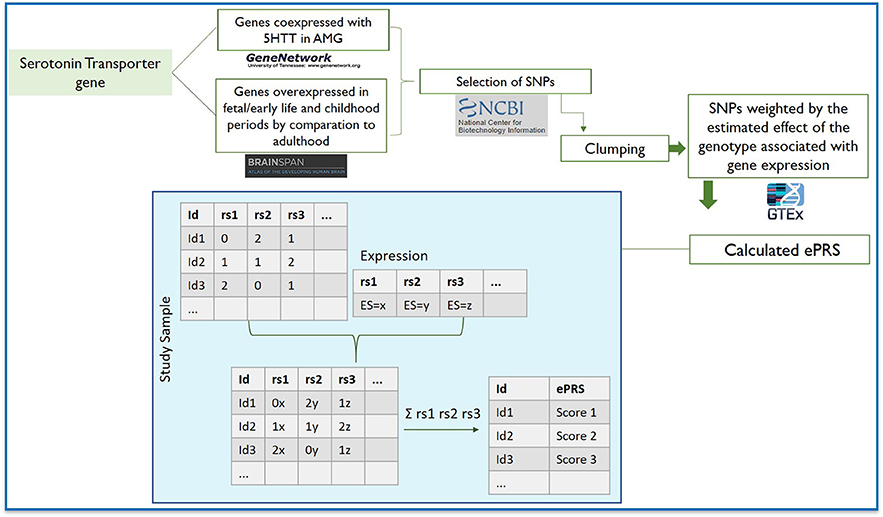
**GeneNetwork**(A) was used to generate a list of genes co-expressed with the serotonin 5-HTT receptor gene in the amygdala in mice (Mulligan et al., 2016).4, retaining only the genes with absolute value of co-expression correlation higher or equal to 0.5.

The gene list generated by **GeneNetwork**(A) was then filtered using **BrainSpan**(B) to identify consensus transcripts enriched in the fetal and childhood human brain (Miller et al., 2014)5. Since we were interested in genes that were active during early developmental periods, we selected autosomal transcripts expressed in the amygdala with at least 1.5-fold more expressed during fetal and child development (all fetal samples and up to the first 5 years of age) as compared to adult samples. The gene network is represented in Figure 1.

Based on their functional annotation in the National Center for Biotechnology Information, U.S. National Library of Medicine, **NCBI Variation Viewer** (C), using GRCh37.p13, we gathered all the SNPs from these genes and merged this list with the SNPs from the **GTEx data**(D) in human amygdala to form a list of common SNPs. The list of common SNPs was then filtered by subjecting it to linkage disequilibrium clumping (r2 < 0.25), which resulted in a total of 463 SNPs.

In ePRS calculation, alleles at a given cis-SNP were weighed by the estimated effect of the genotype on gene expression (expression quantitative trait loci from GTEx, in which the effect allele is the alternative allele). Final ePRS was obtained by summation over all SNPs accounting for the sign of correlation coefficient between the genes and 5HTT gene expression. For more information on ePRS calculation, see Hari Dass et al. (2019)6. The summation of these values from the total number of SNPs provides the amygdala 5-HTT-ePRS score. Figure 2 illustrates the steps involved in the creation of the ePRS.

**Figure 1**: Representation of the gene’s interactions of the amygdala 5-HTT co-expression network. The larger the border, the stronger the relationship of the gene to other genes. The size of the nodes represents the number of incoming relationships with neighbors. The bigger the node, the higher the relationships of other genes to the target gene. The edges represent co-expression. Color and thickness were used to identify the most co-expressed genes, darker, and thicker represent higher co-expression.



**Figure 2**. Flowchart depicting the steps involved in creating the expression-based polygenic risk score based on genes co-expressed with serotonin transporter gene in amygdala using gene co-expression databases: GeneNetwork was used to generate a co-expression matrix with 5-HTT gene in the amygdala in mice (absolute value of the co-expression correlation r ≥ 0.5); BrainSpan was then used to identify consensus human transcripts from this list; BrainSpan was also used for selecting genes differentially expressed at ≥1.5-fold during child and fetal development as compared to adulthood within the same brain areas. Based on their functional annotation in the National Center for Biotechnology Information, U.S. National Library of Medicine, we gathered all the existing SNPs from these genes and subjected this list of SNPs to linkage disequilibrium clumping. We used a count function of the number of alleles at a given SNP (rs1, rs2…) weighted by the estimated effect of the genotype associated with gene expression. The sum of these values from the total number of SNPs provides the amygdala 5-HTT ePRS score.

**Necessary tools and resources**

PRSoS7 :<https://github.com/MeaneyLab/PRSoS>, description and installation instructions provided.

PLINK8 [www.cog-genomics.org/plink/1.9/](http://www.cog-genomics.org/plink/1.9/) , https://www.cog-genomics.org/plink/1.9/general\_usage , description and installation instructions provided.

The script should be executed using Bash command prompt.

**Steps for the use of the tool**

* Download all provided files into one folder
* Edit the path to the PRSoS script in *run\_ePRS\_5HTT.sh* (PRSOS="/**path\_to\_PRSOS\_folder**/PRSoS.py")
* Prepare genotyping data in \*.gen format. Make sure a corresponding sample file is placed along the \*.gen file in the same folder.
* **/ePRS\_score/** subfolder will be created inside the folder where the ePRS script is located. The results will be generated in **/ePRS\_score/** subfolder.
* To run the script simply type: **bash run\_ePRS\_5HTT.sh**

The script will finish processing with the message “Successfully wrote scores to /ePRS\_score/score\_prs.score.csv” and display the time it took to calculate the score.

Files created in the ePRS\_score subfolder are the following:

* *score\_prs.log* is a log file of PRSoS (provided just for information purposes),
* *score\_prs.score.csv* is the result file containing the calculated ePRS, and
* *score\_prs.snplog.csv* is the file containing the rsids of the SNPs included in the ePRS.

**Description of the results**

Description of score\_prs.score.csv file

The file contains three comma-separated columns: ID1 is a subject ID (obtained from the sample file), SNP\_count\_1.0 is a number of SNPs included in the ePRS, PRS\_1.0 is the calculated ePRS.

ID1,SNP\_count\_1.0,PRS\_1.0

P1,602,0.02207469465049833

P2,602,0.006845179335548168

P3,602,0.06661886149501667

Description of score\_prs.snplog.csv file

The file contains three comma-separated columns: PRS\_1.0 is a SNPrsid included in the ePRS, PRS\_1.0\_flag is the effect allele of this a SNP, and Discard shows the SNPs not included in the ePRS score (discarded).

PRS\_1.0,PRS\_1.0\_flag,Discard

ADAMTS7-rs112654305,A1,

ADAMTS7-rs113458918,A1,

ADAMTS7-rs11633351,A1,

**Possible application of the script**

To use for other types of ePRS

The script can also be utilized to calculate ePRS in other tissues and for other gene networks. To implement it, the process of constructing the gene network, gathering all the SNPs, combining with GTEx data and further clumping should be repeated for the specific tissue of interest and the gene network of interest. The resulted data set should replace then the current *gwas\_nodup\_noambig.txt*, which is designed specifically for 5HTT network in amygdala. Important, all columns from the original *gwas\_nodup\_noambig.txt* file should be kept and their order should not be modified.

Note on calculating regular PRS

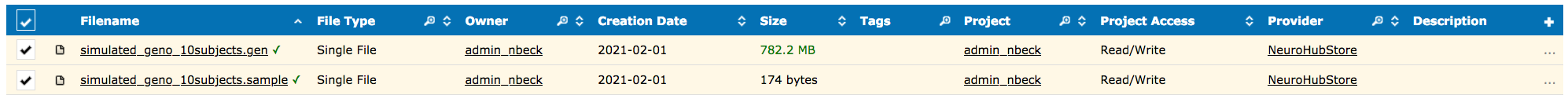
The script can also be used to calculate a regular polygenic risk score (PRS) at a particular threshold providing the GWAS is subsetted to the list of SNPs to be included in the score and which should replace the current *gwas\_nodup\_noambig.txt*. Important, all columns from the original *gwas\_nodup\_noambig.txt* file should be kept and their order should not be modified.

**Run** **5HTT ePRS in CBRAIN/NeuroHub:**

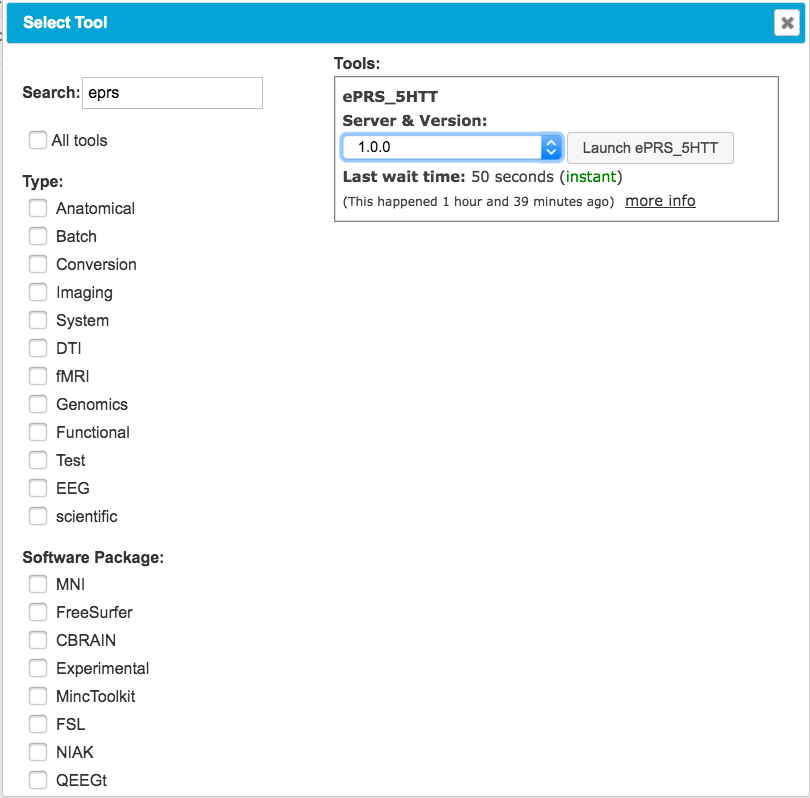
You should have a CBRAIN/NeuroHub account (<https://portal.cbrain.mcgill.ca/>).

Then you should upload your \*.sample and \*.gen files in CBRAIN. To do that you can use the SFTP of CBRAIN (<https://portal.cbrain.mcgill.ca/doc/manual/uploading.html>).

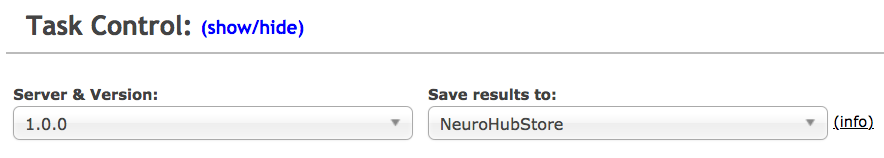
Once your \*.sample file and \*.gen file is in CBRAIN you:

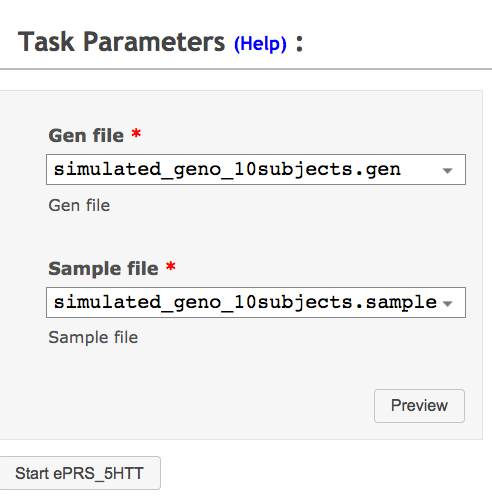
1. Need to select it

2. Search and select for the EPRS\_5HTT tool



3. You will have to select some parameters:





4. Then the task will be setup and then run in CBRAIN:



5. Finally you will be able to access the output\_folder.

**Known issues**

No issues have been identified as of now.

**Versions**

Current version 1.0

**Acknowledgements**

We thank Meaney Lab and Silveira Lab Bioinformatics Core team and past and current staff for the hard work on developing, implementing and updating the tool; we also thank students and collaborators for important feedback over the years.

**References**

1. Silveira PP, Pokhvisneva I, Parent C, et al. Cumulative prenatal exposure to adversity reveals associations with a broad range of neurodevelopmental outcomes that are moderated by a novel, biologically informed polygenetic score based on the serotonin transporter solute carrier family C6, member 4 (SLC6A4) gene expression. *Dev Psychopathol.* 2017;29(5):1601-1617.

2. de Lima RMS, Barth B, Arcego DM, et al. Amygdala 5-HTT Gene Network Moderates the Effects of Postnatal Adversity on Attention Problems: Anatomo-Functional Correlation and Epigenetic Changes. *Frontiers in neuroscience.* 2020;14(198).

3. Miguel PM, Deniz BF, Deckmann I, et al. Prefrontal cortex dysfunction in hypoxic-ischaemic encephalopathy contributes to executive function impairments in rats: Potential contribution for attention-deficit/hyperactivity disorder. *The world journal of biological psychiatry : the official journal of the World Federation of Societies of Biological Psychiatry.* 2017:1-14.

4. Mulligan MK, Mozhui K, Prins P, Williams RW. GeneNetwork: a toolbox for systems genetics. In: *Systems Genetics.* Springer; 2017:75-120.

5. Miller JA, Ding S-L, Sunkin SM, et al. Transcriptional landscape of the prenatal human brain. *Nature.* 2014;508(7495):199-206.

6. Hari Dass SA, McCracken K, Pokhvisneva I, et al. A biologically-informed polygenic score identifies endophenotypes and clinical conditions associated with the insulin receptor function on specific brain regions. *EBioMedicine.* 2019;42:188-202.

7. Chen LM, Yao N, Garg E, et al. PRS-on-Spark (PRSoS): a novel, efficient and flexible approach for generating polygenic risk scores. *BMC Bioinformatics.* 2018;19(1):295.

8. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.